Anatomy segmentation evaluation with sparse ground truth data

Li, Jieyu, Udupa, Jayaram, Tong, Yubing, Wang, Lisheng, Torigian, Drew

Jieyu Li, Jayaram K. Udupa, Yubing Tong, Lisheng Wang, Drew A. Torigian, "Anatomy segmentation evaluation with sparse ground truth data," Proc. SPIE 11313, Medical Imaging 2020: Image Processing, 113131G (10 March 2020); doi: 10.1117/12.2549327

SPIE.

Event: SPIE Medical Imaging, 2020, Houston, Texas, United States

Anatomy segmentation evaluation with sparse ground truth data

Jieyu Li [a,b], Jayaram K. Udupa [b,*], Yubing Tong [b], Lisheng Wang [a], Drew A. Torigian [b]

[a] Institute of Image Processing and Pattern Recognition, Department of Automation, Shanghai Jiao Tong University, 800 Dongchuan RD, Shanghai, 200240, China; [b] Medical Image Processing Group, Department of Radiology, University of Pennsylvania, 602 Goddard building, 3710 Hamilton Walk, Philadelphia, PA, 19104, United States

ABSTRACT

The performance and evaluation of segmentation algorithms will benefit from large fully annotated data sets, but the heavy workload of manual contouring is unrealistic in clinical and research practice. In this work, we propose a method of automatically creating pseudo ground truth (p-GT) segmentations of anatomical objects from given sparse manually annotated slices and utilize them to evaluate actual segmentations. Sparse slices are selected spatially evenly on the whole slice range of the target object, where one slice is selected to conduct manual annotation and the next *t* slices are skipped, repeating this process starting from one end of the object to its other end. A shape-based interpolation (SI) strategy and an object-specific 2D U-net based deep learning (DL) strategy are investigated to create p-GT. The largest *t* value where the created p-GT is considered to be not statistically significantly different from the actual ground with its natural imprecision due to variability in manually specified ground truth is determined as the optimal *t* for the considered object. Experiments are conducted on ~300 computed tomography (CT) studies involving two objects – cervical esophagus and mandible and two segmentation evaluation metrics – Dice Coefficient and average symmetric boundary distance. Results show that the DL strategy overwhelmingly outperforms the SI strategy, where ~95% and ~66-83% of manual workload can be reduced without sacrificing evaluation accuracy compared to actual ground truth data via the DL and SI strategies respectively. Furthermore, the p-GT with optimal *t* is able to evaluate actual segmentations with accurate metric values.

Keywords: medical image segmentation, segmentation evaluation, sparse ground truth, deep learning

1. INTRODUCTION

Numerous 3D anatomy segmentation methods have emerged since the advent of tomographic imaging modalities in the 1970s. Early methods were purely image-based[1, 2] which needed ground truth (or reference) segmentations only for segmentation evaluation since they did not harvest anatomic priors from existing data sets. Although such approaches continue to seek new frontiers, methods that exploit priors in various forms have emerged during the past 2-3 decades and have shown significant gain in segmentation robustness and accuracy. These later methods may be generically referred to as model-based since they employ some form of model to encode prior information such as anatomic and geographic models[3], atlases[4], deep neural networks[5], etc. However, for these methods, fully annotated ground truth (GT) segmentations that capture the very variability over a human population of focus they purport to encode is of fundamental importance, some of them requiring a large number of such data sets for robust model building alone, not to mention for evaluation as well.

There are two main issues with generating GT segmentations. (I1) Expense: GT reference is most typically provided by manual (human expert) contouring of anatomical objects in medical imaging. Thus, generating fully manually large GT sets becomes impractical and expensive[6]. (I2) Imprecision: Owing to various reasons such as human subjectivity in interpreting boundaries in images, lack of standard ways of defining objects, or variations in the interpretation of (pseudo) standards (when they exist), GT segmentations have imprecision. The magnitude of these imprecisions is object-specific. Small, non-compact, and sparse objects entail larger degrees of imprecision proportionate to their size compared to large,

* jay@pennmedicine.upenn.edu

Medical Imaging 2020: Image Processing, edited by Ivana Išgum, Bennett A. Landman,
Proc. of SPIE Vol. 11313, 113131G · © 2020 SPIE · CCC
code: 1605-7422/20/$21 · doi: 10.1117/12.2549327

well-defined, and compact objects. I1 is purely a cost issue. I2, however, raises several conceptual issues. Although both I1 and I2 have been examined to some extent in the literature, this area calls for a lot more attention in view of the promises suggested by deep neural network models. Most importantly, the practical question of the savings that result in cost as a function of the imprecision in GT data as a result of its "sparsification" has not been examined so far. In other words, is it feasible to simulate full GT segmentations from sparse GT data such that the simulated GT is as good (or imperfect) as, but not worse than, the GT generated by human experts? The cost saving then will depend on the degree of sparsity affordable for the sparse GT data and will be tied with the second issue I2.

As to I1, several algorithms have been proposed in the past for object-matching or model-training via partially annotated samples[7, 8], but the lack of fully annotated samples may raise confusion in segmentation evaluation, showing the disconnection with I2. Ref[9] has investigated methods to evaluate the quality of crowdsourced non-expert annotations and to fuse crowdsourced labels to help in biomedical segmentation research. The Expectation-Maximization-based STAPLE[10] method and its extensions are a series of methods commonly used to generate consensus GT from multiple manual segmentations. None of these efforts addressed I1 and I2 jointly or one as a function of the other.

In this paper, keeping I1 in mind, we investigate two pseudo GT (p-GT) generation strategies. They both start from manual annotations given on only a sparse set of slices among all slices occupied by the target object and then fill in GT segmentations automatically for those skipped slices. We then bring in I2 by investigating how the generated p-GT would vary, as we change the degree of sparseness, in comparison to the imprecision existing in the actual GT. The degree of sparseness at which the deviation of the p-GT (generated by the chosen p-GT strategy) with respect to the actual GT is as good (or as bad) as the imprecision in the actual GT will then be considered as the optimal (highest) affordable sparseness level (and hence cost saving) for that p-GT strategy. The p-GT generation strategy can then be considered to behave like an alternative to an expert segmenter or a pseudo expert segmenter. To investigate how much manual workload can be reduced when maintaining human-level imperfection, experiments are conducted on different degrees of sparseness for slice selection for the two proposed p-GT strategies. The performance of p-GT in segmentation evaluation is verified via evaluation of actual segmentations in comparison to evaluation by actual fully annotated GT.

2. METHODS

The proposed p-GT generation approach, called SparseGT, is represented in Figure 1. There are two key aspects to the SparseGT method: (i) sparse slice selection strategy, where we first manually create GT segmentations on sparsely selected slices, and (ii) segmentation filling strategy, where we fill slices not selected for GT segmentation with pseudo segmentations. Both operations are performed in an object-specific manner. Finally, we use the p-GT data to evaluate segmentations of the same object samples in comparison to actual (full) GT data.

2.1 Sparse slice selection

The method of selecting sparse slices is illustrated in Figure 2. Let t denote the *degree of sparseness*, with the idea being that higher values of t indicate higher degrees of sparseness. The idea is to select one slice and then skip the next t slices and repeat this process starting from one end of the object up to its other end in the direction orthogonal to the slice plane. Manual contouring (segmentation) is then conducted only on the sparsely selected slices. Subsequently, the contours of the object on the skipped slices are filled by an appropriate p-GT strategy.

Determination of a proper (optimal) t value for each object is crucial in the SparseGT method: Large values of t would greatly reduce the workload of expert segmenters but may increase distinguishable deviation from the actual ground truth. Let N_O be the smallest number of slices covering an object O (such as the mandible) over a population S of patient images that include O. An extreme t with least manual work load would be $t = \lfloor (N_O - 2)/1 \rfloor$, where only the slices on two ends are selected and manually contoured and the remaining slices are skipped, or $t = \lfloor (N_O - 3)/2 \rfloor$, where the middle slice and the slices at the two ends are manually contoured and all other slices in between are skipped. More generally, if we select $n \leq N_O$ slices with roughly uniform spacing, $t = \lfloor (N_O - n)/(n-1) \rfloor$. The proper choice for t is a value that is large enough to reduce manual workload, while simultaneously producing a p-GT that does not deviate significantly from actual ground truth compared with the deviation found among expert segmenters. The selection of t is object-specific and also depends on the strategy employed for creating p-GT from sparse data.

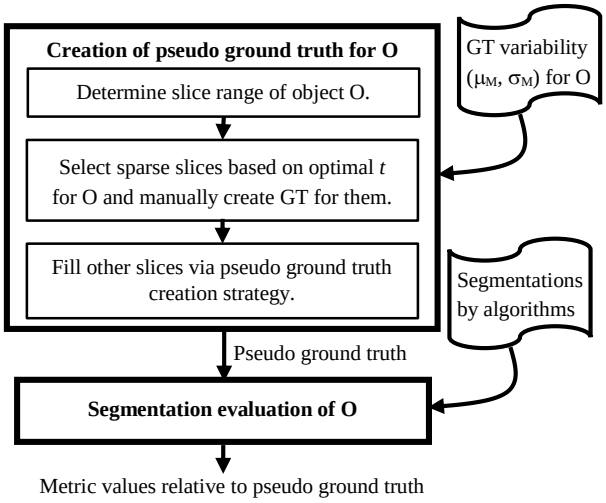

Figure 1. A schematic representation of the SparseGT method.

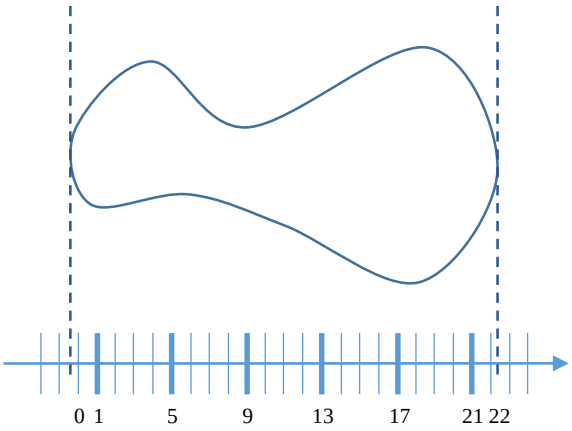

Figure 2. Illustration of selecting positions of sparse slices for the given GT data. Manual contouring is performed on every $(t + 1)^{th}$ slice. In the figure, $t = 3$, $N = 23$, and $n = 6$. The bold vertical lines indicate the selected sparse slices.

2.2 Segmentation filling for slices not selected for GT contouring

The manually annotated contours on the selected sparse slices implicitly represent a sampling of the manner in which the expert segmenter behaves in the contouring task. The achievable maximum sparseness for slice selection is not only object-specific, but also depends on the ability of the p-GT strategy to emulate the behavior of the expert segmenter. Two segmentation filling strategies are investigated in this work: one is a straight-forward strategy of shape-based interpolation (SI) and the other is deep-learning (DL) based. Shape-based interpolation, first proposed in Ref[11], is a family of approaches wherein the object is estimated at in-between slices by following the slice-to-slice shape of the object. The method reported in Ref[11] first applies a 2D distance transform to the binary shape in each given slice with the convention that distances inside the object are positive and those outside are negative. To estimate a slice in-between two given slices, the distance values are interpolated and then the sign of the estimated distance value is used to determine if a pixel is inside or outside the object (see Figure 3(a)).

For the DL-strategy, a 2D U-net[12] based network is used for creating p-GT in this paper; any other suitable architecture can be utilized as well. The network is illustrated in Figure 3(b). *Input*: A $6t + 8$ channel image which contains 4 continuous sparse slices with the associated manual contours together with blank slices corresponding to the skipped slices in between and the corresponding original intensity image slices at the same spatial positions. *Output*: A t channel binary mask predicted for the central block of t skipped slices. Taking Figure 2 as an example where $t = 3$ is selected as the sparseness parameter, the target object in this sample occupies $N = 23$ image slices numbered 0 to 22, and $n = 6$ slices – 1, 5, 9, 13, 17 and 21 – are selected and contoured manually while other slices are skipped. When we intend to automatically create contours on slices between Nos. 5 and 9, i.e., on slices numbered 6, 7, and 8, we create a 26-channel input image constituted by binary slices and original intensity slices numbered 1 to 13 where only foreground on slices numbered 1, 5, and 9 are marked as 1 and all other regions and all other slices are marked as background 0 on the binary slices. After all blocks of skipped slices are filled with predicted binary masks, sparse slices with manual contours and blocks of skipped slices with predicted contours are concatenated following the original sequence and restored to the original spatial positions. This restored binary image is called *DL-p-GT*. Analogously, we refer to the binary images created by shape-based interpolation as *SI-p-GT*. These pseudo ground truth segmentations can then be used as an alternative to fully manually annotated GT data in evaluating segmentations produced by an algorithm.

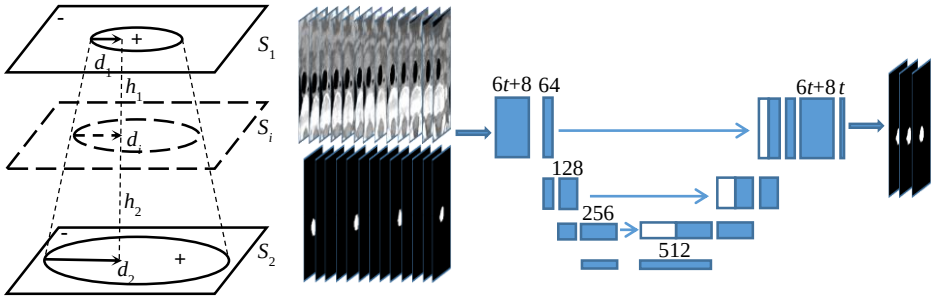

Figure 3. Illustration of segmentation filling strategies. (a) Shape-based interpolation approach. (b) 2D U-net based deep learning method.

2.3 Inter-segmenter variation and estimation of the optimum value of *t*

Individual differences among human segmenters are hard to eliminate. Even clinical professionals like dosimetrists in Radiation Oncology departments, who manually contour organs at risk for radiation therapy (RT) planning, show considerable inter-segmenter variation. Figure 4 presents examples of two objects in the Head & Neck body region as they appear in CT images: cervical esophagus, which is a thin tube-like sparse object with low boundary contrast, and mandible which is a less challenging object with both sparse and non-sparse aspects and mostly good boundary contrast. s_1 and s_2 denote two expert dosimetrists who separately conducted manual annotation on the same object samples.

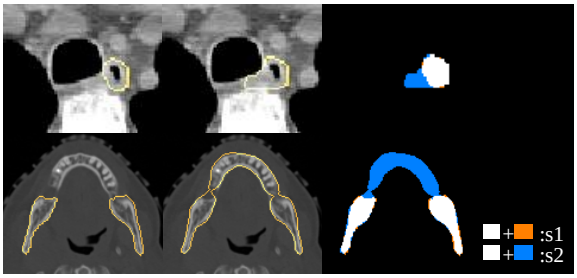

Figure 4. Illustration of inter-segmenter differences: Cervical esophagus (top row) and mandible (bottom row). Substantial differences can be seen between the two segmentations, where white regions stand for agreements by two segmenters, and orange and blue regions stand for inter-segmenter differences.

Perfect GT does not exist as inter-segmenter differences always exist. The SparseGT method exploits this fact and tries to generate p-GT that is as good as the actual GT data available to us with all of their imperfections. Firstly, we select a segmentation evaluation metric. Here, we demonstrate the approach by employing two most commonly used metrics, namely Dice Coefficient (DC) and average symmetric boundary distance (ASD). For each object O whose segmentations by an algorithm A are to be evaluated, we then obtain the variability of these metric values among a group G of expert segmenters by having them create GT segmentations of O on the same given set S of images. Typically, S required for the SparseGT approach is much smaller than the size of the data sets required for training model-based segmentation algorithms. More importantly, this variability needs to be established only once for O. In this paper, we employ two experts to establish this variability for demonstration purposes. For each metric M, we describe its variability by a pair (μ_M, σ_M), where μ_M denotes the mean value and σ_M denotes the standard deviation of M over all samples of S among all combinations of expert segmenters in G taken two at a time, where one is taken as the reference segmentation with respect to which the other expert's segmentations are evaluated via M. Analogously, we determine the variability (μ_{Mp}, σ_{Mp}) of the pseudo

ground truth generated by taking different experts in G into account, where μ_{Mp} denotes the mean value of M and σ_{Mp} is the standard deviation of M over all experts in G. In this paper, for demonstration purposes, we have considered only one expert in G. We then determine the optimum value t_O of the degree of sparseness as the largest value of t where the deviation (μ_{Mp}, σ_{Mp}) from p-GT becomes statistically indistinguishable from the variability (μ_M, σ_M) among experts in the actual full GT. Once t_O is determined for an object, the p-GT generated for O using t_O can be used for evaluating the performance of algorithm A in segmenting O.

3. EXPERIMENTS, RESULTS, AND DISCUSSION

3.1 Experiments

This retrospective study was conducted following approval from the Institutional Review Board at the Hospital of the University of Pennsylvania along with a Health Insurance Portability and Accountability Act waiver. Experiments are conducted on computed tomography (CT) images of the Head & Neck body region focusing on two objects – cervical esophagus (CtEs) and mandible (Mnd). The objects are chosen to represent different shape and size characteristics and different degrees of challenges for segmentation. A set of 298 3D images with full ground truth for both objects are used in our experiments. The GT data constitute real clinical data as contoured by a dosimetrist for the routine RT planning of patients with Head & Neck carcinoma. For 81 data sets among the above cohort, two dosimetrists fully annotated contours of considered objects expressly for the purpose of recording the natural imprecision that exists in manual GT in this application domain, and thus we also have reference GT variation data as well in our annotations, while we have annotations from only one expert segmenter different from the other two for other images in the entire set.

The data sets with two tracings from two dosimetrists are separately divided into training and test samples, where about 20% of the samples are randomly selected and compose the test set. The remaining samples form the training set to determine t_O, which is verified on the test set to check if the test set can also yield indistinguishable deviation with respect to expert variability. The voxel size in our data sets varies from $0.93 \times 0.93 \times 1.5$ mm^3 to $1.6 \times 1.6 \times 3$ mm^3. The object sizes are also variable among subjects, where object samples of different subjects occupy varying numbers of slices, 18-71 slices for CtEs and 17-86 slices for Mnd. The bounding box size is also determined for each object based on its largest occupied range and to fit the input size expected by the DL network. The region of interest (ROI) size of samples is set in multiples of 8, since there are three convolutional layers with stride 2 in both U-net based networks – 64×96 for CtEs and 160×144 for Mnd.

In our experiments, we assume that the extreme (ideal) case is one where the middle slice and two end slices are used for sparse selection among N_O slices occupied by the considered object, which corresponds to $t_3 = \lfloor (N_O - 3)/2 \rfloor$, and for the more general cases, the degree of sparseness t_n is selected such that $1 \leq t_n \leq t_3$, where $n = \lfloor (N_O - 1)/(t_n + 1) \rfloor + 1$ slices are selected. The best sparseness factors t_O are determined by using the training set. Then, the determined sparseness factors are evaluated on a separate test set which also should not yield distinguishable deviations with respect to the reference GT. The SI strategy does not need a training stage in generating p-GT, while for the DL-based strategy, 2-fold cross validation is conducted to generate p-GT for the training set based on different t_n and the models for the test set are trained on the whole training set.

The validity of p-GT in segmentation evaluation is demonstrated by comparing metric values by p-GT with values evaluated by actual full GT. Root mean squared error (RMSE) is used to evaluate the deviation of metric values measured by p-GT from the values measured from actual GT. RMSE ε is calculated as shown below, where α stands for one of the metrics DC and ASD, $\mathcal{I}_b = \{I_{1,b}, \ldots, I_{N,b}\}$ stands for binary masks of manually created full ground truth, $\mathcal{J}_b = \{J_{1,b}, \ldots, J_{N,b}\}$ denotes binary masks generated by segmentation algorithms, and $\mathcal{P}_b = \{P_{1,b}, \ldots, P_{N,b}\}$ represents the created p-GT. ε is a function of the sparseness parameter t under different p-GT creation strategies and for different objects. Segmentation algorithm A considered to be evaluated is the AAR-RT method described in Ref[13].

$$\varepsilon = \sqrt{\frac{1}{N}([\alpha(I_{1b}, J_{1b}) - \alpha(P_{1b}, J_{1b})]^2 + ,..., + [\alpha(I_{Nb}, J_{Nb}) - \alpha(P_{Nb}, J_{Nb})]^2)}$$

3.2 Results and discussions

(1) Determining optimal sparseness t_O

Image examples for objects CtEs and Mnd are shown in Figure 5 and the optimum sparseness factor t_O are determined as illustrated separately in the left columns of Figures 6 and 7. The optimal sparseness factor t_O is determined as the largest t where the metric values is statistically indistinguishable from the same metric that shows the variability in actual full GT. Since we utilize DC and ASD as metrics to measure the difference between segmentations from different human segmenters and also between manual and pseudo GT, the estimated t_O should receive agreement from the two metrics. T-test is conducted to determine statistical significance with *p-value* < 0.05. The cases with significant differences are marked 'o' in the left column of Figures 6 and 7, while cases with insignificant differences are marked '*'. Quantitative results are listed in Tables 1 and 2, and insignificant cases are marked in bold.

We can infer from plots and Tables 1 and 2 that, for cervical esophagus, t_O = 5 for the SI strategy and t_O = 14 for the DL strategy, and for mandible, t_O = 2 for the SI strategy and t_O = 16 for the DL strategy. The reduction of manual workload can be estimated as 1 / (t + 1) of original workload, where only one slice out of t + 1 slices needs manual contouring and the remaining t slices can be skipped and filled automatically via SI or DL strategies. That means, by using the straight forward shape-based interpolation strategy without any training, the manual workload can be reduced to 16.7% of the full workload for cervical esophagus and to 33.3% for mandible. Furthermore, given a proper set of training samples, with the DL strategy, the workload can be further reduced to 6.67% and 5.88% for CtEs and Mnd, respectively, of the full manual workload, leading to 93.33% and 94.12% savings in workload for the two objects, respectively!

We should note that, among the population of 298 samples, t_3 = 16 is the ideal sparseness factor can be reached for most of the samples of CtEs and Mnd. Even with the maximum sparseness, the error of DL-p-GT for Mnd can still be lower than the inter-segmenter difference, which shows a strength of the DL strategy in that, with human-level accurate recognition and proper and consistent object definition (a fundamental tenet of the AAR-RT approach[13]), it is able to yield high-quality delineation masks for the mandible. The excellent performance on the mandible may be also attributed to the lower level of challenge in its segmentation. If we compare plots for CtEs and Mnd, although the performance on the training set and the test set of CtEs show similar tendency with increasing t and *p-value* > 0.05 in most cases between cross-validation results of the training set and results of the test set, the performances for mandible are similar for both metrics, which means CtEs is a more variable object compared to Mnd. The optimal t_O value for the SI strategy reflects how regular the object shape is along the axis orthogonal to the scanning plane. In this sense, we may infer that although Mnd is less challenging than CtEs in segmentation, it is less regular and has more shape change from slice to slice.

Table 1. Difference of pseudo ground truth of cervical esophagus with respect to actual manual ground truth for different sparseness factors. Mean and standard deviation values are listed.

O = CtEs	DC				ASD (mm)			
	DL-p-GT		SI-p-GT		DL-p-GT		SI-p-GT	
	$S_{training}$	S_{test}	$S_{training}$	S_{test}	$S_{training}$	S_{test}	$S_{training}$	S_{test}
Inter-segmenter	0.88 0.04				0.38 0.28			
t = 4	0.93 0.02	0.93 0.02	0.92 0.03	0.91 0.04	0.18 0.05	0.18 0.05	0.21 0.06	0.23 0.08
t = 5	0.92 0.02	0.92 0.02	0.89 0.05	**0.88** **0.05**	0.21 0.09	0.2 0.06	0.29 0.11	**0.31** **0.12**
t = 6	0.92 0.02	0.91 0.03	**0.87** **0.06**	0.86 0.07	0.23 0.07	0.23 0.07	**0.38** **0.18**	**0.41** **0.19**
t = 7	-		0.84 0.08	0.82 0.09	-		0.49 0.27	0.55 0.29
t = 8	0.9 0.02	0.9 0.03	0.81 0.1	0.78 0.12	0.28 0.09	0.26 0.1	0.62 0.36	0.68 0.41

t = 12	**0.88** **0.03**	**0.89** **0.03**	0.69 0.13	0.66 0.16	**0.36** **0.19**	**0.33** **0.11**	1.19 0.65	1.34 0.76
t = 14	**0.88** **0.03**	**0.88** **0.04**	0.63 0.14	0.62 0.15	**0.4** **0.14**	**0.39** **0.2**	1.57 0.83	1.66 0.87
t = 15	**0.87** **0.03**	**0.87** **0.03**	0.6 0.15	0.6 0.15	0.45 0.19	**0.43** **0.18**	1.83 0.96	1.81 0.92
t = 16	0.86 0.04	**0.87** **0.04**	0.57 0.15	0.6 0.15	0.47 0.23	**0.43** **0.22**	2.03 1.06	1.82 1.05

Table 2. Difference of pseudo ground truth of mandible with respect to actual manual ground truth for different sparseness factors. Mean and standard deviation values are listed.

O = Mnd	DC				ASD			
	DL-p-GT		SI-p-GT		DL-p-GT		SI-p-GT	
	$S_{training}$	S_{test}	$S_{training}$	S_{test}	$S_{training}$	S_{test}	$S_{training}$	S_{test}
inter-segmenter	0.91 0.02				0.35 0.11			
t = 2	0.97 0.01	0.97 0.01	0.94 0.02	0.94 0.02	0.12 0.05	0.11 0.05	0.26 0.09	0.24 0.08
t = 3	0.96 0.01	0.96 0.02	**0.91** **0.03**	**0.91** **0.03**	0.15 0.05	0.14 0.06	0.44 0.16	0.43 0.16
t = 4	0.96 0.01	0.96 0.02	0.88 0.04	0.88 0.04	0.17 0.06	0.16 0.09	0.61 0.23	0.61 0.28
t = 8	0.94 0.02	0.94 0.02	0.75 0.09	0.74 0.1	0.23 0.09	0.22 0.09	1.48 0.7	1.54 0.8
t = 12	0.94 0.02	0.94 0.02	0.58 0.15	0.57 0.16	0.25 0.09	0.24 0.11	3.03 1.77	3.16 2.2
t = 16	0.93 0.02	0.93 0.02	0.36 0.13	0.34 0.14	0.29 0.12	0.26 0.1	6.35 3.39	6.54 3.39

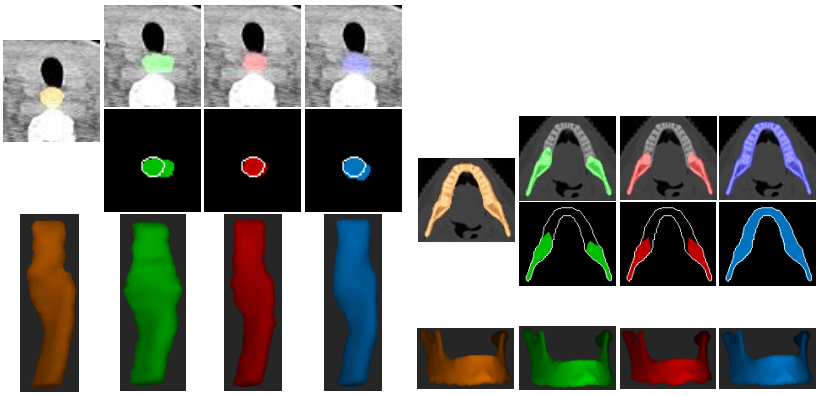

Figure 5. Image examples for CtEs and Mnd. Manual ground truth (m-GT) are generated by two expert segmenters – s1 and s2. Pseudo ground truth (p-GT) are generated from sparse m-GT by s1 and the optimal t_O via SI and DL strategies respectively. 2D binary masks are overlaid on the intensity images and overlaid by s1 contours, and the corresponding surface models are presented as well.

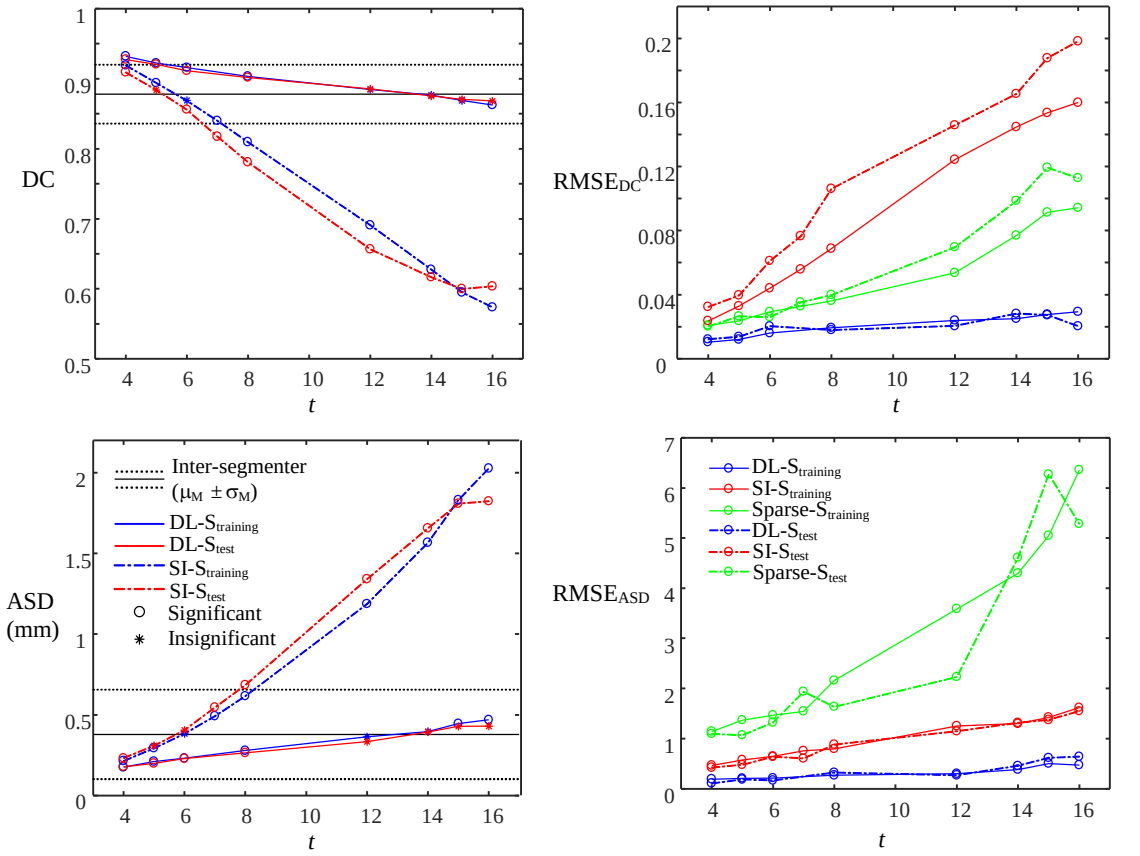

Figure 6. Illustrations of experimental results for different t for cervical esophagus CtEs. Left column: Variation of p-GT taking actual GT from one expert as reference. The optimum sparseness factor t_O is determined by the largest t without yielding statistically significant difference compared to (μ_M, σ_M). Right column: Root mean squared error ε of evaluation metric values of actual segmentations via AAR-RT method based on DL- and SI-p-GT strategies.

(2) Segmentation evaluation with pseudo ground truth

Illustrations for p-GT for CtEs and Mnd in evaluation of actual segmentations are shown in the right columns of Figures 6 and 7, and quantitative results are listed in Tables 3 and 4, respectively. We observe from the plots and also from quantitative results that DL-p-GT shows best capability to replace manual ground truth in that it generates least error in evaluation measures compared to simply sparse or SI-p-GT. Also, for DL-p-GT, with increasing sparseness, i.e., increasing t, the error does not increase as rapidly as the other two kinds of pseudo ground truth.

When comparing the RMSEs of p-GT with optimal t separately based on DL- and SI- strategies, we found that the yielded errors are actually similar, which means that the p-GT sets created by the two strategies for t_O will both have similar acceptable evaluation measures with only slight deviation. The RMSEs of p-GT also show the influence the inter-segmenter difference may have on segmentation evaluation. Specific to the practical usage for segmenting CtEs and Mnd, if the dataset for training or model building and the test dataset are contoured by different expert segmenters, there will be an error of 0.03 and 0.02 in DC or an error of 0.6 and 0.2 mm in ASD. This error may be blamed on inter-segmenter differences but not on the real capability of the trained model or the algorithm. Inter-segmenter differences also show upper bounds for how accurate the automatic segmentation algorithms can become, while beyond those bounds it is doubtful, if directly verified on other sources of data sets, that the algorithms will be able to yield segmentations with as good evaluation measures as with the training data sets. With the explosive development of deep learning architectures, we believe that there are several algorithms that are able to reach or surpass this upper bound for objects like Mnd, while there is still room for sparse and challenging objects like CtEs to improve[14, 15].

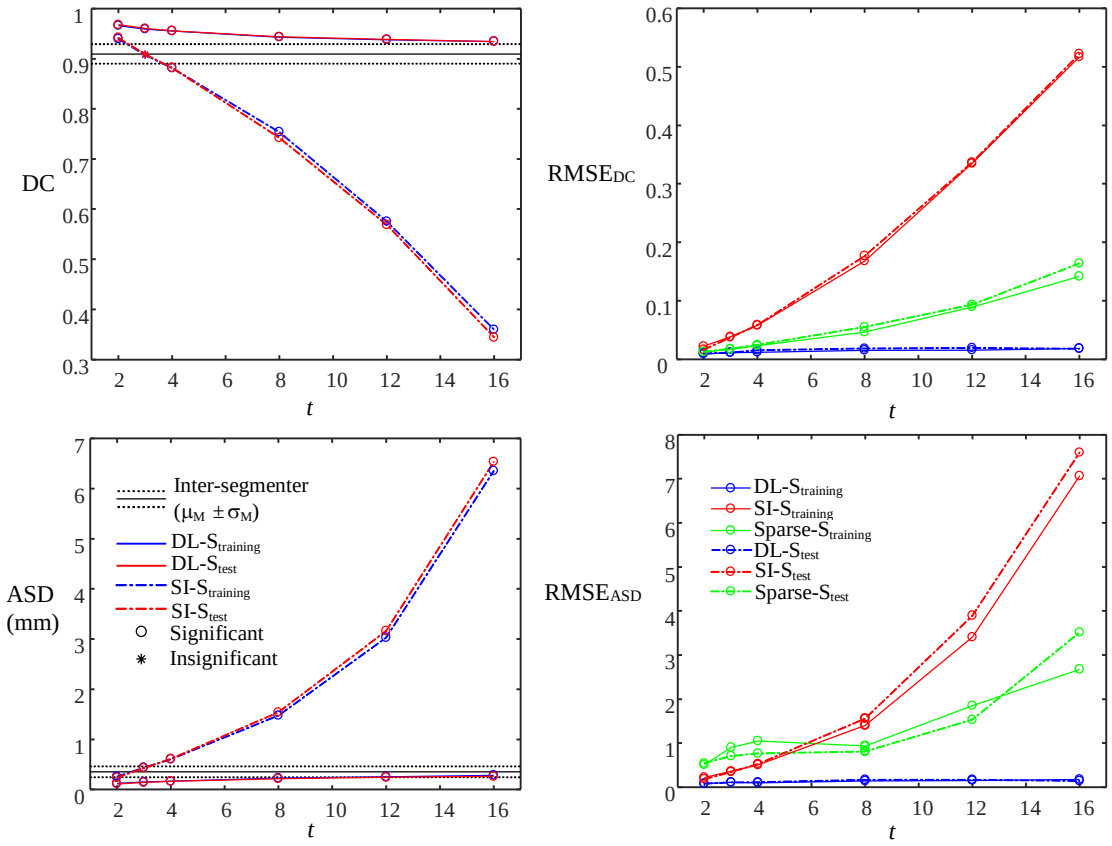

Figure 7. Illustrations of experimental results for different t for mandible Mnd. Notations are same as in Figure 6.

Table 3. Root mean squared error (RMSE) of pseudo ground truth of cervical esophagus on actual segmentation evaluation compared to by actual manual ground truth.

O = CtEs	DC						ASD					
	DL-p-GT		Sparse-GT		SI-p-GT		DL-p-GT		Sparse-GT		SI-p-GT	
	$S_{training}$	S_{test}	$S_{training}$	S_{test}	$S_{training}$	S_{test}	$S_{training}$	S_{test}	$S_{training}$	S_{test}	$S_{training}$	S_{test}
t = 4	0.01	0.01	0.02	0.02	0.02	0.03	0.19	0.11	1.15	1.1	0.47	0.42
t = 5	0.01	0.01	0.02	0.03	0.03	0.04	0.21	0.19	1.37	1.07	0.58	0.48
t = 6	0.02	0.02	0.03	0.03	0.04	0.06	0.22	0.17	1.47	1.33	0.65	0.64
t = 7	-		0.03	0.04	0.06	0.08	-		1.55	1.94	0.76	0.61
t = 8	0.02	0.02	0.04	0.04	0.07	0.11	0.27	0.32	2.16	1.64	0.8	0.88
t = 12	0.02	0.02	0.05	0.07	0.12	0.15	0.3	0.27	3.59	2.23	1.25	1.15
t = 14	0.02	0.03	0.08	0.1	0.14	0.17	0.39	0.46	4.3	4.61	1.3	1.32
t = 15	0.03	0.03	0.09	0.12	0.15	0.19	0.51	0.62	5.05	6.27	1.42	1.37
t = 16	0.03	0.02	0.09	0.11	0.16	0.2	0.47	0.64	6.36	5.28	1.62	1.55

Table 4. Root mean squared error (RMSE) of pseudo ground truth of mandible on actual segmentation evaluation compared to by actual manual ground truth.

O = Mnd	DC						ASD					
	DL-p-GT		Sparse-GT		SI-p-GT		DL-p-GT		Sparse-GT		SI-p-GT	
	$S_{training}$	S_{test}	$S_{training}$	S_{test}	$S_{training}$	S_{test}	$S_{training}$	S_{test}	$S_{training}$	S_{test}	$S_{training}$	S_{test}
t = 2	0.01	0.01	0.01	0.01	0.02	0.02	0.08	0.08	0.51	0.54	0.22	0.19

$t = 3$	0.01	0.01	0.02	0.02	0.04	0.04	0.11	0.11	0.9	0.71	0.36	0.35
$t = 4$	0.01	0.02	0.02	0.02	0.06	0.06	0.1	0.11	1.05	0.77	0.51	0.53
$t = 8$	0.02	0.02	0.05	0.06	0.17	0.18	0.14	0.17	0.94	0.81	1.4	1.56
$t = 12$	0.02	0.02	0.09	0.09	0.33	0.34	0.15	0.17	1.85	1.53	3.41	3.9
$t = 16$	0.02	0.02	0.14	0.16	0.52	0.52	0.18	0.15	2.68	3.51	7.07	7.6

4. CONCLUSIONS

In this paper, our goal was to addresses a gap that currently exists in segmentation evaluation, namely, to seek an answer to the question "Is it possible to create machine-generated ground truth which is just as good as the full manual ground truth from sparse human annotated data sets?" Recognizing the fact that human-drawn ground truth will never be perfect, we investigated a novel method named SparseGT of creating pseudo ground truth vastly more efficiently. With a fraction of the manual workload needed for creating full ground truth, we have shown that the created pseudo ground truth works at least as well as the full ground truth in terms of accuracy. No such work currently exists. We presented two automated and object-specific strategies – shape-based interpolation (SI) and deep learning (DL) – for creating pseudo ground truth from sparse ground truth data sets. Two objects, cervical esophagus and mandible in the Head & Neck body region, with different shapes, sizes, and segmentation challenges have been investigated utilizing ~300 CT data sets. Two segmentation evaluation metrics are studied – DC and ASD, and the maximum sparseness factor which yields consonant indistinguishable differences measured by both metrics with respect to the imprecision that exists in actual manual ground truth is determined as the optimal sparseness factor. The DL method performs overwhelmingly better than the SI strategy. We have demonstrated that ~95% of manual workload can be alleviated via the DL strategy without sacrificing accuracy compared to actual ground truth data. Even via the SI strategy, which is a straight forward method that does not need any model-training, the workload can be reduced by ~66-83%. As such, it can serve as a potential method to enlarge data sets for deep learning training, if not directly used for generating pseudo ground truth.

We are further investigating the underlying core ideas presented in this work in several directions with the inclusion of: larger data sets, all major organs in the Head & Neck body region, other body regions and their organs, non-uniform and shape-dependent sparse slice selection, etc.

REFERENCES

[1] Liu, H. K., "Two- and three-dimensional boundary detection," Comput. Graph. Image Processing 6, 123–134 (1977).
[2] Herman, G. T., Srihari, S., and Udupa, J., "Detection of Changing Boundaries in Two- and Three-Dimensions," Proceedings of the Workshop on Time Varying Imagery, (eds.) N.I. Badler, J.K. Aggarwal, University of Pennsylvania, Philadelphia, Pennsylvania 14-16, (1979).
[3] Udupa, J. K., Odhner, D., Zhao, L., Tong, Y., Matsumoto, M. M., Ciesielski, K. C., Falcao, A. X., Vaideeswaran, P., Ciesielski, V., Saboury, B., and Mohammadianrasanani, S., "Body-wide hierarchical fuzzy modeling, recognition, and delineation of anatomy in medical images," Med. Image Anal. 18(5), 752-771 (2014).
[4] Shi, C., Cheng, Y., Wang, J., Wang, Y., Mori, K., and Tamura, S., "Low-rank and sparse decomposition based shape model and probabilistic atlas for automatic pathological organ segmentation," Med. Image Anal. 38, 30-49 (2017).
[5] Drozdzal, M., Chartrand, G., Vorontsov, E., Shakeri, M., Jorio, L. D., Tang, A., Romero, A., Bengio, Y., Pal, C., and Kadoury, S., "Learning normalized inputs for iterative estimation in medical image segmentation," Med. Image Anal. 44, 1–13 (2018).
[6] Schipaanboord, B., Boukerroui, D., Peressutti, D., van Soest, J., Lustberg, T., Kadir, T., Dekker, A., van Elmpt, W., and Gooding, M., "Can atlas-based auto-segmentation ever be perfect? Insights from extreme value theory," IEEE Trans. Med. Imaging 38(1), 99-106 (2018).
[7] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O., "3D U-Net: learning dense volumetric segmentation from sparse annotation," MICCAI, 424-432 (2016).
[8] Koch, L. M., Rajchl, M., Bai, W., Baumgartner, C. F., Tong, T., Passerat-Palmbach, J., and Rueckert, D., "Multi-atlas segmentation using partially annotated data: methods and annotation strategies," IEEE Trans. Pattern Anal. Mach. Intell. 40(7), 1683-1696 (2017).

[9] Gurari, D., Theriault, D., Sameki, M., Isenberg, B., Pham, T. A., Purwada, A., Zhang, C., Wong, J. Y. and Betke, M. "How to collect segmentations for biomedical images? A benchmark evaluating the performance of experts, crowdsourced non-experts, and algorithms," In 2015 IEEE winter conference on applications of computer vision, 1169-1176 (2015).
[10] Warfield, S. K., Zou, K. H., and Wells, W. M. "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," IEEE Trans. Med. Imaging 23(7), 903 (2004).
[11] Raya, S.P. and Udupa, J.K., "Shape-based interpolation of multidimensional objects," IEEE Trans. Med. Imaging. 9(1), 32-42 (1990).
[12] Ronneberger, O., Fischer, P. and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," MICCAI, 234-241 (2015).
[13] Wu, X., Udupa, J.K., Tong, Y., Odhner, D., Pednekar, G.V., Simone II, C.B., McLaughlin, D., Apinorasethkul, C., Lukens, J., Mihailidis, C., Shammo, G., James, P., Tiwari, A., Wojtowicz, L., Camaratta, J. and Torigian, D.A., "AAR-RT – A system for auto-contouring organs at risk on CT images for radiation therapy planning: principles, design, and large-scale evaluation on head-and-neck and thoracic cancer cases," Med. Image Anal. 54, 45-62 (2019).
[14] Chan, J.W., Kearney, V., Haaf, S., Wu, S., Bogdanov, M., Reddick, M., Dixit N., Sudhyadhom A., Chen J., Yom S.S. and Solberg T.D., "A convolutional neural network algorithm for automatic segmentation of head and neck organs at risk using deep lifelong learning," Med. Phys. 46(5), 2204-2213 (2019).
[15] Tong, N., Gou, S., Yang, S., Cao, M., Sheng, K., "Shape constrained fully convolutional DenseNet with adversarial training for multiorgan segmentation on head and neck CT and low-field MR images," Med. Phys. 46(6), 2669-2682 (2019).